Supplementary Material: Strategyproof Mean Estimation from Multiple-Choice Questions

1. Proof of Theorem 3

To prove the theorem, we reduce from the following problem. Given a rational $x \in [0,1]$ and nonnegative integer weights $\alpha_1, \ldots, \alpha_n$, WEIGHTED-BINOMIAL-MEDIAN (WBM) asks for a median of the random variable

$$Z := \sum_{i=1}^{n} \alpha_i \, \mathsf{Bernoulli}(x),$$

where the $\mathsf{Bernoulli}(x)$ are independent (and identically distributed).

This weighted binomial distribution (WBD) is comparable to the Poisson binomial distribution (PBD) in that both generalize the binomial distribution. However the PBD is an unweighted sum of Bernoulli random variables with distinct probabilities x_i, while the WBD is a sum of Bernoulli random variables with a common x but distinct integer weights.

Lemma 1. WBM *is $\#\mathcal{P}$-Hard.*

Proof. In order to show to show that WBM is $\#\mathcal{P}$-complete, we will reduce from the counting version of the knapsack problem, which is known to be $\#\mathcal{P}$-complete (**?**): Given a list of nonnegative integer weights $w_1, \ldots, w_n$ and an integer capacity W, #Knapsack asks how many sets $S \subseteq [n]$ exist such that $\sum_{i \in S} w_i \leq W$. And we will make use of a slight variant of counting knapsack: Given an integer k, a list of nonnegative integer weights $w_1, \ldots, w_n$ an integer capacity W, and an integer threshold N, $k\#$Knapsack finds $|\mathcal{S}|$, where

$$\mathcal{S} := \{S \subseteq [n] : \sum_{i \in S} w_i \leq W \text{ and } |S| = k\}.$$

It can be seen that $k\#$Knapsack is $\#\mathcal{P}$-complete via an easy reduction from #Knapsack: given an instance of #Knapsack, simply query $k\#$Knapsack for all values of k and return the sum of the answers.

Turning to the hardness of WBM, we begin by arguing that WBM may be assumed to return the largest possible median. This is because, for an instance of WBM given by $(x, \alpha_1, \ldots, \alpha_n)$, we may instead take a perturbed probability $\bar{x} = x + \gamma$. By choosing γ small enough, we can ensure that the median $\bar{m}$ of $\bar{Z} := \sum_i \alpha_i \mathsf{Bernoulli}(\bar{x})$ is a median of Z, but that it is the largest possible such median. Informally, we may tweak x gently enough that we preserve the median but break any median ties.

Formally, let F_Z be the cumulative density function (CDF) of Z. Since Z is a distribution comprised solely of atoms of weight $x^k(1-x)^{n-k}$ for $k \in [n]$, it suffices to find some perturbation γ for which

$$F_Z(m) - F_{\bar{Z}}(m) < a,$$

where m is a median of Z and a is a lower bound on the size of an atom in both Z and $\bar{Z}$. To show that we may choose such an a, note we may assume that $(1-x)^n \leq 1/2$, since otherwise the largest possible m is 0, and similarly that $\bar{x}^n \leq 1/2$, since otherwise we may easily check if the largest possible m is $\sum_{i \in [n]} \alpha_i$. Among all Z for which $x^n \leq 1/2$ and $(1-x)^n \leq 1/2$, the smallest possible atom is of size $\frac{1}{2}(2^{1/n} - 1)^n$, and so $a := 1/n^n$ is a lower bound on the atom size in Z for any value of x that concerns us.

Since Z is atomic, we then have that

$$F_Z(y) = \sum_{z \leq y} \Pr[Z = z] \tag{1}$$

$$= \sum_{S \subseteq [n]} x^{|S|}(1-x)^{n-|S|} \, \mathbb{1}_{\{\sum_{i \in S} w_i \leq y\}} \tag{2}$$

and so

$$\frac{\partial F_Z(y)}{\partial x} \leq \sum_{S \subseteq [n]} \frac{\partial}{\partial x} x^{|S|}(1-x)^{n-|S|} \leq n2^n. \tag{3}$$

Therefore taking $\gamma = \frac{a}{n2^n}$ will suffice, and $\bar{x} = x + \gamma$ will have a binary representation which is polynomial in the number of input bits.

We now reduce from $k\#$Knapsack. Given an instance of $k\#$Knapsack described by $(k, w_1, \ldots w_n, W)$, let $\Gamma := \langle k \rangle + \sum_i \langle w_i \rangle + \langle W \rangle$ be the length of the binary representation of these integers. For each i, let

$$\alpha_i := G + w_i,$$

where $G := (n+1)\sum_i w_i$. If $Z = \sum_i \alpha_i \mathsf{Bernoulli}(x)$ for some rational $x \in [0,1]$, then since the w_i are positive, the support of Z is clustered to the left of the integers $0, G, \ldots, nG$. Specifically, we have by Equation (2) that

$$\begin{aligned} F_Z(Gk) &= \sum_{S \subseteq [n]} x^{|S|}(1-x)^{n-|S|} \, \mathbb{1}_{\{\sum_{i \in S} w_i \leq Gk\}} \\ &= \sum_{j=0}^{k-1} \binom{n}{j} x^j (1-x)^{n-j}, \end{aligned}$$

and so $F_Z(Gk)$ can be computed in time polynomial in $\Gamma + \langle x \rangle$.

Next, with k given, consider a binary search over (rational) x which searches for the largest possible x for which $m \leq Gk + W$. Once the binary search is far enough along and the change in x is sufficiently small, $F_Z(m)$ approaches $1/2$ and the remaining change possible in $F_Z(m)$ will be small with respect to the atomic lower bound a. We may terminate our search, say, when $F_Z(m) \in [1/2, 1/2 + a/10]$. At this point m is the largest value of size at most $Gk + W$ in the support of Z, and so by this maximality of $m \leq Gk + W$,

$$\begin{aligned} F_Z(m) &= \sum_{S \subseteq [n]} x^{|S|}(1-x)^{n-|S|} \, \mathbb{1}_{\{\sum_{i \in S} w_i \leq m\}} \\ &= \sum_{j=0}^{k-1} \binom{n}{j} x^j (1-x)^{n-j} + |\mathcal{S}_k| \, x^k (1-x)^{n-k}. \end{aligned}$$

At this point a is much smaller than the other terms, and we may solve for $|\mathcal{S}_k|$, round, and solve $k\#$Knapsack:

$$|\mathcal{S}_k| \in \frac{1/2 \pm a/10 - \sum_{j=0}^{k-1} \binom{n}{j} x^j (1-x)^{n-j}}{x^k (1-x)^{n-k}}$$

It remains only to justify that this binary search for x terminates sufficiently quickly. By Equation (3) in order to guarantee that $F_Z(m)$ is within $a/10$ of $1/2$ it suffices to guarantee that the binary search step for x has size at most $\frac{a}{10n2^n}$. This requires $\log\left(10n^{n+1}2^n\right)$ steps, which is polynomial in n. □

Proof of Theorem 3. We reduce from WBM. If $k = 1$ then the reduction is immediate: if each of the P_i is a scaled down copy of $\alpha_i \mathsf{Bernoulli}(x)$, then finding the optimal report for the random variable $\sum_i P_i$ amounts to finding the (scaled down) median of $\sum_i \alpha_i \mathsf{Bernoulli}(x)$.

More generally, given an instance of WBM described by $(x, \alpha_1, \ldots, \alpha_n)$, we will construct an instance of our problem, MAE-ESTIMATOR, for any $k \geq 2$ for which determining optimal partitions and reporting scheme will solve our instance of WBM.

Our P_i will be discrete distributions given by

$$\Pr\left[p_i = \frac{1}{2k}\right] = \frac{1-x}{k} \tag{4}$$

$$\Pr\left[p_i = \frac{1+\delta\frac{\alpha_i}{\sum_t \alpha_t}}{2k}\right] = \frac{x}{k} \tag{5}$$

$$\Pr\left[p_i = \frac{2j-1}{2k}\right] = \frac{1}{k} \qquad \text{for } j = 2, \ldots, k\,. \tag{6}$$

We will choose δ small enough such that the optimal partition of each of the P_i necessarily groups the atoms described in Equation (4) and Equation (5) together, and gives each of the atoms of Equation (6) its own interval in the partition. To find such a δ, first consider the "good" case when the partitions are of this form. In this case, there are k^n total boxes, each with weight $1/k^n$. Within each box C, the distribution of ℓ_1 norms has range upper bounded by $\delta/(2k)$. Within each C, the range of this distribution is an upper bound on the ℓ_1 distance between any atom in C and the optimal report for C. Therefore, a loose upper bound on total MAE is

$$\sum_{c \in [k]^n} P(C_c)\frac{\delta}{2k} = \frac{\delta}{2k}. \tag{7}$$

On the other hand, consider the "bad" case when at least one of the partitions groups either two of the Equation (6) atoms together or the Equation (5) atom together with at least one of the Equation (6) atoms. Assume without loss of generality that the $i = 1$ partitioning is "bad". We will focus on the case when an Equation (5) and at least one Equation (6) atom are grouped together (because it is an interval, necessarily $j = 2$ is included), since in the best case it is the least costly scenario. Because of the product structure of the boxes induced by the partitions, for every pair of vectors u and u' in the support of P of the form

$$u = \left(\frac{1+\delta\frac{\alpha_i}{\sum_j \alpha_j}}{2k}, u^-\right) \qquad u' = \left(\frac{3}{2k}, u^-\right),$$

where $u^- \sim \prod_{j=2}^n P_j$, necessarily u and u' are contained in the same box. Therefore among each pair of u and u', at least $M_{u^-} = \frac{\min\{x,1-x\}}{k}\prod_{j=2}^n P_j(u_j^-)$ mass must travel $\|u'\|_1 - \|u\|_1$ to the estimate for their shared box, which yields a lower bound on the error of

$$\sum_{u^-}\left(\frac{1-\delta}{k}M_{u^-}\right) = \frac{(1-\delta)\min\{x,1-x\}}{k^2}. \tag{8}$$

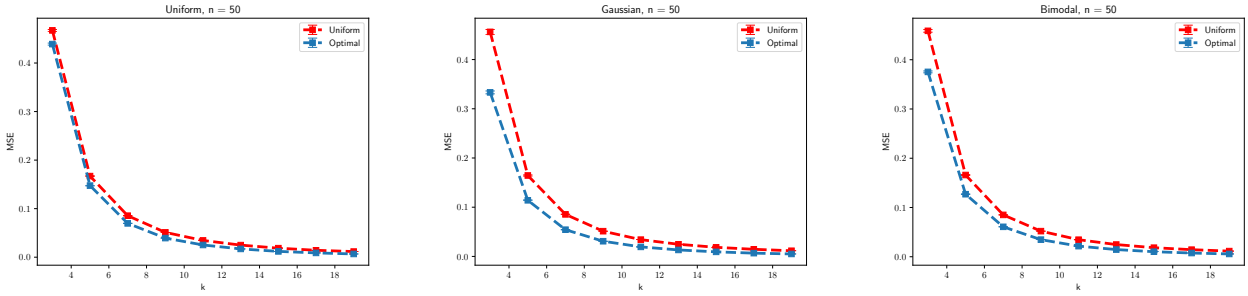
By Equations (7) and (8), choosing a $\delta < \frac{\min\{x,1-x\}}{k}$ guarantees that the optimal partitioning for our instance is the "good" partitioning, and so all of the Equation (4) and Equation (5) atoms appear in the same box $C^\star := \prod_i B_{i,1}$.

Recall that by **??**, the MAE-minimizing estimate for a fixed box C is a median of the distribution of ℓ_1 norms of the vectors $u \in C$ according to P. Therefore MAE-ESTIMATOR finds some MAE-optimal report $r^\star$ for the box $C^\star$, which by Equation (4) and Equation (5) implies that $\frac{r^\star - n/2k}{\delta}$ is a median of $\sum_i \alpha_i \mathsf{Bernoulli}(x)$, solving the given instance of WBM. □

2. Additional Experiments

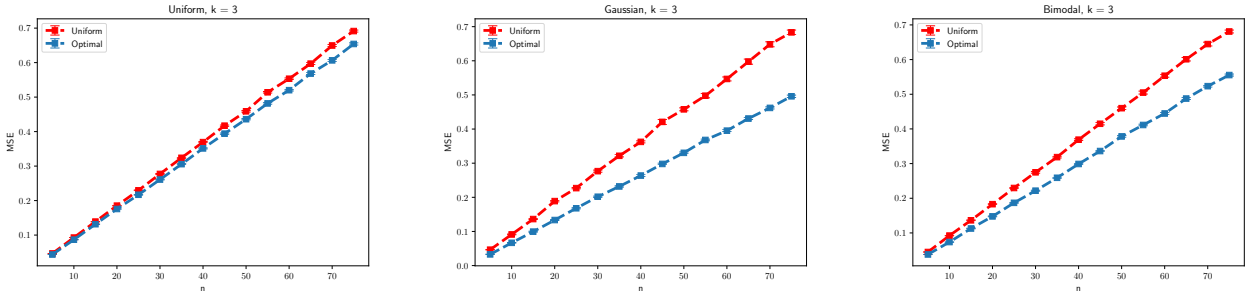
Here we give a more thorough account of the experimental performance of the MSE-optimal estimation scheme as compared to the uniform estimation scheme, benchmarked against the families of distributions described in **??**.

Figure 1 shows their relative performance for a fixed value of $n = 50$ and a range of possible k, with constant sample distribution support of size 100. On the other hand, Figure 2 shows their relative performance for a fixed value of $k = 3$ and a range of possible n, again with constant sample distribution support of size 100.



(a) Uniform (b) Gaussian (c) Bimodal

Figure 1: MSE of the uniform and optimal algorithms for fixed $n = 50$ and a range of k, averaged over 100 distributions sampled from the various families. Bars are standard error of the mean.



(a) Uniform (b) Gaussian (c) Bimodal

Figure 2: MSE of the uniform and optimal algorithms for fixed $k = 3$ and a range of n, averaged over 100 distributions sampled from various families. Bars are standard error of the mean.